DOCUMENT RESUME

ED 401 301 TM 025 850

AUTHOR Stocking, Martha L.
TITLE Revising Answers to Items in Computerized Adaptive Tests: A Comparison of Three Models.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-96-12
PUB DATE Apr 96
NOTE 40p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; Cognitive Processes; Comparative Analysis; *Computer Assisted Testing; *Error Correction; *Mathematical Models; Simulation; Test Bias; Testing Problems; *Test Items
IDENTIFIERS *Answer Changing (Tests); Large Scale Assessment; Revision Processes

ABSTRACT

The interest in the application of large-scale computerized adaptive testing has served to focus attention on issues that arise when theoretical advances are made operational. Some of these issues stem less from changes in testing conditions and more from changes in testing paradigms. One such issue is that of the order in which questions are answered within a test or a separately timed test section. In linear testing, this order of responses is entirely under the control of the test-taker, who can omit questions, look ahead at questions, and return and revise answers to previous questions. The attempt to permit the same, or even reasonably restricted, control in adaptive testing can unintentionally result in transferring to the test-taker control over which items are chosen for administration, threatening both test fairness and accuracy. This paper explores, using simulations, three models of permitting test-taker control over revising previous answers in the context of adaptive testing. Even under a worst-case model of test-taker revising behavior, two of the models of permitting item revisions work well in preserving test fairness and accuracy and one model studied may also preserve some cognitive processing styles developed by test-takers for a linear testing environment. (Contains 4 figures, 5 tables, and 23 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-96-12

ED 401 301

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

RESEARCH REPORT

REVISING ANSWERS TO ITEMS IN COMPUTERIZED ADAPTIVE TESTS: A COMPARISON OF THREE MODELS

Martha L. Stocking

Educational Testing Service
Princeton, New Jersey
April 1996

TM025850

BEST COPY AVAILABLE

REVISING ANSWERS TO ITEMS IN COMPUTERIZED ADAPTIVE TESTS: A COMPARISON OF THREE MODELS[1,2]

Martha L. Stocking

Educational Testing Service
Princeton, New Jersey 08541

February, 1996

[1] This research was supported in part by the Program Research Planning Council of Educational Testing Service.

[2] The author wishes to thank Marion Horta, Charlotte Kuh, Charlie Lewis, Craig Mills, Bob Mislevy, Nancy Petersen, Manfred Steffen, Marilyn Wingersky, and Michael Zieky for their time, patience, and insights.

Copyright © 1996. Educational Testing Service. All rights reserved.

REVISING ANSWERS TO ITEMS IN COMPUTERIZED ADAPTIVE TESTS: A COMPARISON OF THREE MODELS

Abstract

The interest in the application of large-scale computerized adaptive testing has served to focus attention on issues that arise when theoretical advances are made operational. Some of these issues stem less from changes in testing conditions and more from changes in testing paradigms. One such issue is that of the order in which questions are answered within a test or separately timed test section. In linear testing, this order of responses is entirely under the control of the test-taker, who can omit questions, look ahead at questions, and return and revise answers to previous questions. The attempt to permit the same, or even reasonably restricted, control in adaptive testing can unintentionally result in transferring to the test-taker control over which items are chosen for administration, threatening both test fairness and accuracy. This paper explores, using simulations, three models of permitting restricted test-taker control over revising previous answers in the context of adaptive testing. Even under a worst-case model of test-taker revising behavior, two of the models of permitting item revisions work well in preserving test fairness and accuracy and one model studied may also preserve some cognitive processing styles developed by test-takers for a linear testing environment.

Key Words: computerized adaptive testing, revising answers to items, test-taking strategies, response ordering.

REVISING ANSWERS TO ITEMS IN COMPUTERIZED ADAPTIVE TESTS: A COMPARISON OF THREE MODELS

Introduction

Recent advances in psychometrics and computing technology have led to the development of a testing paradigm that is very different from linear paper-and-pencil testing -- computerized adaptive testing (CAT; see, for example, Eignor, Way, Stocking & Steffen, 1993; Lord, 1977; Schaeffer, Steffen & Golub-Smith, 1993; Stocking & Swanson, 1993; Wainer, Dorans, Flaugher, Green & Mislevy, 1990; Weiss, 1982). As interest in large-scale implementation of modern adaptive testing has increased, particularly for high-stakes testing programs (Jacobson, 1993), increasing attention has been focussed on issues that arise when theoretical advances are made operational (see, for example, Mills & Stocking, 1995).

Some of these issues stem less from changes in testing conditions and more from changes in testing paradigms. An example of such an issue is that of the order in which responses to questions are given within a test or a separately timed section of a test. In linear paper-and-pencil testing, candidates take a single form (or parallel forms) of a test, and while the test designers determine the collection of questions and arrange them in a certain order, the actual order of responding to those questions is determined by the test-taker. Thus test-takers may skip or omit questions that may be too hard, and return to them after they have responded to other questions. Or test-takers may engage in extended cognitive processing that might lead them to revise answers previously given. Much advice on good test-taking strategies is predicated on the assumption that response ordering is under the

control of test-takers, within the constraints of time limits. This advice is appropriate for the linear testing paradigm.

Adaptive tests are tests in which items are selected, one at a time, from a large pool of items in such a fashion as to be appropriate for a test-taker (the test "adapts" to the test-taker). Typically, the next item to be administered is selected based on the answers given to all previous items, and if these responses tend to be incorrect, easier items are chosen, whereas if these responses tend to be correct, harder items are chosen. Because the items are chosen dynamically as the test is administered, certain features of the linear paper-and-pencil testing context do not transfer easily to this new context. In particular, because the set of items to be administered to a particular test-taker is not specified in advance of test administration, it is difficult, and perhaps impossible, to devise methods that maintain the identical test-taker control over the actual order in which the questions are answered. Moreover, some good test-taking strategies that depend upon test-taker control of response ordering may be less appropriate in the context of the adaptive testing paradigm.

For example, it is probably not possible to permit test-takers to omit and return to an item. To do so would decrease the measurement efficacy of the adaptive test, and reduce the efficiency of the test design. But more importantly, permitting a test-taker to omit items raises issues of test fairness since it allows test-takers to review an entire item pool, possibly leading to widespread pool compromise. Fortunately, it is probably less important to provide for omits/skips in the CAT context since, in theory, test-takers are rarely presented with items that are too hard for them -- a feature of the adaptive test design that is not present in linear testing.

On the surface, it looks equally unlikely that provisions can be made in CAT to permit the revision of previous answers. Certainly such provisions are likely to decrease the efficiency of the test design since the selection of items can become less than optimal. For example, suppose a test-taker answered the first 10 items presented, and then decided to revise their answer to the fifth item. The selection of items six through 10 based on the original response to the fifth item is probably suboptimal when compared to what would have been chosen for items six through 10 based on the revised answer to item 5.

More importantly, Wainer (1992) points out that permitting revisions to previous items has the potential to threaten adaptive test fairness in that through this mechanism it may be possible for test-takers to construct an inappropriately easy test on which they score very well. Wainer suggests that a test-taker should intentionally respond incorrectly, if possible, to any item presented until the end of the test is reached. At the end of the test the test-taker has designed the easiest possible test from the item pool, and can now revise as many answers as possible to correct. This strategy seems clearly in the best interests of test-takers who are naturally motivated to achieve the highest possible test-score.

This strategy is likely to work to the benefit of test-takers regardless of the underlying psychometric model upon which test scores are based and regardless of the particular method of estimating test-taker proficiency, for example, either maximum likelihood (Lord, 1980) or Bayes modal estimates (Mislevy, 1986). This is because the administration of an inappropriate test to a test-taker results in larger errors in the estimate of test-taker proficiency (ignoring any possible bias), and low to middle ability test-

takers in particular are likely to benefit from less precise proficiency estimates.

The attempt to permit the same control over the order of item administration in CAT as in linear tests can unintentionally result in giving control to the test-taker over the actual items administered. If all test-takers took advantage of this feature, CAT would not be unfair to any test-taker, but would be worthless as a measuring instrument from the perspective of test-score users. If only some (even only a few) test-takers took advantage of this feature, then CAT would become unfair to those test-takers who did not capitalize on this strategy. Nevertheless, it could be argued that it may be desirable from the cognitive processing perspective alone to provide CAT test-takers some facility for revising previous answers.

Both Lunz, Bergstrom & Wright (1992) and Stone & Lunz (1994) have studied the effect of item revisions on the psychometric properties of CAT when actually administered to real test-takers. The results of both studies contrast with the Wainer speculation. Lunz, Bergstrom & Wright studied examinees taking a variable-length adaptive certification test for practice (that is, scores did not count) where the test stopped when test scores could be bounded away from the cut score with 90% confidence. Revision of items after the test was completed resulted in a slight decrease (1%) in test efficiency. On average, post-revision ability estimates were slightly higher, however, only 1% of the pass/fail decisions were altered by such revisions.

Stone & Lunz studied similar variable-length adaptive tests, although these tests apparently were not practice tests. Similar results were found -- a slight decrease in test efficiency, a slight increase in post-revision ability estimates, and a small (6%) change in pass/fail decisions.

The small effects reported in both studies were due, in part, to the low rates of item revisions reported. In the Lunz, Bergstrom & Wright study, the average test length was 96 items and the average number of responses altered was 2, or approximately 2% of the items administered. In the Stone & Lunz study, two variable-length adaptive tests were studied, both with minimum test lengths of 50 items and maximum test lengths of 100 items. Those test-takers who received 50-item tests revised about 6% of the items administered. Those taking Test 1 who received more than 50 items had test lengths that averaged 81 items and revised about 4% of the items administered; comparable Test 2 examinees received 85 items on average and revised about 5% of them.

Both sets of authors acknowledge that long strings of response changes in one direction, for example, from wrong to right, can have profound impact on test efficiency as well as test scores. The fact that such changes were not observed in the CATs studied may indicate that test-takers were not sufficiently well-informed of the strategies outlined by Wainer that could be used to improve their test scores in the adaptive testing environment but not in the linear testing environment.

If it is desirable to provide some test-taker control over item response ordering in the environment of high-stakes, well-coached adaptive testing, it may be possible to do so by employing mechanisms that are different from those used in the linear testing context. This paper explores, using simulated data, the consequences of three models for permitting item revisions in adaptive testing. The next section presents information about revising behavior in the framework of linear testing. The following section describes the type of adaptive testing employed in the simulation experiments. Subsequent sections describe the actual adaptive tests used in the simulation

experiments and present a worst-case model of revising behavior that underlies the experiments and is even more dire than that suggested by Wainer. Finally, the simulation experiments conducted using three different models for permitting item revisions in CAT are described and the results presented. The consequences of these three different models are discussed in terms of test fairness and accuracy.

Revising Behavior in Linear Testing

Little is known about revising behavior in linear paper and pencil testing because no mechanisms exist for capturing such information from answer sheets. Such baseline data from the situation in which motivated test-takers operating in a high-stakes linear testing environment control item ordering may provide useful information, particularly about item revisions that may result from cognitive processing demands.

As part of a multistep study to determine the comparability of linear paper and pencil tests and adaptive testing for the purpose of admission to graduate schools, the Graduate Records Examination Board and Educational Testing Service collected information electronically from 6,977 test-takers who took linear computer-based Quantitative, Verbal, and Analytical Reasoning measures (Schaeffer et al., 1993). The test-takers were well motivated in that their test scores counted for admissions purposes. And since these linear tests were administered by computer, it was possible to capture a complete record of test-taker behavior.

The computer system developed for item presentation allowed test-takers to review previously administered items, supply answers for omitted items and change answers, mark items for later review, look at items that were not yet

administered, and so forth. Every effort was made to provide the same types of facilities, at least functionally, that are available in paper and pencil test administrations of these measures.

Table 1 displays global information about reviewing and revising behavior for each measure. The average number of items answered varied from 97% of the intended test length for the Analytical Reasoning measure, to 99% for the Verbal measure.

Table 1: Overall Reviewing and Revising by Measure

	Quantitative Reasoning	Verbal Reasoning	Analytical Reasoning
Number of items in measure	60	76	50
Average number of items responded to	59.23	75.34	48.54
Average maximum number of items reviewed	8.09	17.73	6.84
Average number of items revised	3.52	8.21	4.84

Table 2: Average Number of Items Revised for Quantitative Measure

Item Response	Final: Omit	Right	Wrong	Total
Initial: Omit	X	1.37	1.31	2.68
Right	.00	X	.13	.13
Wrong	.00	.44	.27	.71
Total	.00	1.81	1.71	3.52

Table 3: Average Number of Items Revised for Verbal Measure

Item Response	Final: Omit	Right	Wrong	Total
Initial: Omit	X	3.35	2.68	6.03
Right	.00	X	.43	.43
Wrong	.00	1.02	.73	1.75
Total	.00	4.37	3.84	8.21

Table 4: Average Number of Items Revised for Analytical Measure

Item Response	Final: Omit	Right	Wrong	Total
Initial: Omit	X	2.22	1.89	4.11
Right	.00	X	.09	.09
Wrong	.00	.43	.21	.64
Total	.00	2.65	2.19	4.84

The item presentation system permitted two different methods of reviewing items. The most explicit method required test-takers to "mark" items for later review, and then permitted return to those items from different positions in the test while skipping any intervening items. The less explicit method permitted test-takers to simply scroll backward or forward through the test. In this less explicit method, test-takers could intentionally stop and review items, perhaps on their way to a particular item, or they could simply pass through items until they reached the item sought. It was possible, of course, to detect 'marked' items and the return to them. However, it was not possible to distinguish intentional as opposed to unintentional review of items if test-takers were scrolling through the items. Therefore, the combination of all types of revisits to items is reported as the 'average maximum number of items reviewed'. On average, the maximum number of items reviewed by test-takers ranged from about 13% to 14% of the total test length for Quantitative and Analytical Reasoning to about 23% for Verbal Reasoning.

Not all revisits to items resulted in changes to responses. The final row in Table 1 reports the average number of items for which the final response given was different from the initial response. Initial responses

include omitted responses. On average, test-takers changed from 6% to 11% of their initial responses.

Tables 2, 3, and 4 show more detailed information about the average number of items revised for the Quantitative, Verbal, and Analytical Reasoning measures. Each row in a table contains results for final response conditional on initial response. The row and column totals add up to the figures presented in the final row of Table 1. An 'X' is used to mark those cells that result in no change and are therefore empty. For example, an initial response of omit and a final response of omit is considered as no change in response, regardless of what may have occurred in between.

It is noteworthy that over 70% of the items revised, for each measure, were items that were originally omitted. The majority of these omissions were discrete items for which test-takers were presumably unsure of the correct response -- the items were too hard for them. However, for each measure, just over 50% of these items were finally answered correctly. The second largest category of items initially omitted were items associated with common stimulus material such as items associated with reading comprehension passages (not shown in table). Evidently it is not uncommon for a test-taker to peruse all items associated with a stimulus before finalizing the response to any item.

Results more strictly comparable to the Lunz, Bergstrom & Wright study and the Stone & Lunz study discussed earlier, where initial omits were not permitted, are shown in the 2 x 2 sub-tables enclosed in dashed lines. For the Quantitative Reasoning measure, there were .84 (.13 + .44 + .27) items revised that were not initially omitted, or about 1% (.84/59.23) of the average number of items administered. For the Verbal Reasoning measure, this figure was about 3%, and for the Analytical Reasoning measure it was about 2%.

About 60% of the initially incorrect responses (.44/.71 for Quantitative, 1.02/1.75 for Verbal, and .43/.64 for Analytical) were changed to final correct responses. This finding, that is, that over half of initial wrong responses that are subsequently revised are changed to final correct responses, is similar to that of the previous studies.

The total rates of item revisions when compared to total test length (excluding omits), that is, the 1%, 3%, and 2% for the Quantitative, Verbal, and Analytical measures are about what was found in both previous studies, although perhaps on the low side. This is presumably because in the previous studies, initial omits were not allowed, therefore some part of the revision behavior previously reported is a consequence of forced initial choices that test-takers might have omitted if they had been permitted to do so.

These data suggest the following conclusions:

1) The majority of item revisions in a linear test come from initial omissions, presumably because the test-taker was confronted with items that were too hard for them. This is an appropriate test-taking strategy in linear testing, but less appropriate in adaptive testing.

2) There is some suggestion that a strategy followed with some frequency in linear testing was to take an holistic approach to items associated with common stimuli -- examine more than one item in the set before finalizing responses to any of them. This method of structuring cognitive processing seems efficient and reasonable.

3) The similarity between these results for linear testing and the results previously reported for adaptive testing strengthen the assertion that in the previous adaptive testing studies, test-takers were not aware of or did not

choose to use the strategies outlined by Wainer for maximizing their scores in the adaptive test environment.

Thus these data are directly informative about revising behavior in linear testing where test-takers can control the order of item administration. However, they, as well as the Lunz, Bergstrom & Wright, and the Stone & Lunz results are only indirectly informative about test-taker revising behavior in the context in which test-takers can use mechanisms for item ordering to also control which items are selected for administration, as is the potential in adaptive testing.

Adaptive Testing With the Weighted Deviations Model

As noted by Davey & Parshall (1995) high-stakes adaptive testing has at least three goals: 1) to maximize test efficiency by selecting the most appropriate items for a test-taker, 2) to assure that the tests measure the same composite of multiple traits for each test-taker by controlling the nonstatistical characteristics of items, such as content, included in the test, and 3) to protect the security of the item pool by controlling the rates at which items can be administered. These goals often compete with one another.

Different approaches to each of these goals yield different algorithms for adaptive testing. The particular algorithm used in this paper is the Weighted Deviations Model (WDM) developed by Swanson & Stocking (1993) and applied to adaptive testing by Stocking & Swanson (1993). This paradigm is characterized by flexible approaches to all three goals of adaptive testing.

In general, any CAT algorithm implicitly orders the items in the pool in terms of their desirability for selection as the next item. Differences in

ordering typically reflect particular definitions of item optimality and particular methods of estimating test-taker ability. Any attempt to control the exposure of items can then be viewed as modifications imposed on this ordering.

In the WDM the item pool is ordered by employing a methodology from the decision sciences that models the behavior of expert test specialists. The WDM ordering explicitly takes into account nonstatistical item properties or features along with the statistical properties of items. This is to insure that each adaptive test produced from a pool matches a set of test specifications and is therefore as parallel as possible to any other test produced from that pool in terms of content and type of items, while being tailored to an individual test-taker in terms of measurement appropriateness. The desired balance between measurement and construct concerns is reflected by the weights given to them, which are chosen by the test designer. The WDM approach also allows specification of overlapping items that may not be administered in the same adaptive test. In addition, it is possible to restrict item selection to blocks of items, either because they are associated with a common stimulus or common directions or any other feature that test specialists deem important. Thus as each item is selected for a test-taker using the WDM, the pool or an appropriate subset of the pool is ordered from most desirable (smallest weighted deviations from desirable test properties) to least desirable (largest weighted deviations from desirable test properties).

In summary, in the WDM, the next item selected for administration is the item that simultaneously

1) is the most appropriate available at a test-taker's estimated ability level, while

2) contributing as much as possible to the satisfaction of all other test specification constraints.

At the same time, it is required that the next item selected for administration

3) does not appear in an overlap group containing an item already administered, and

4) is in the current block (if the previous item was in a block), starts a new block, or is in no block.

In the particular version of the WDM used in this paper, the measure of the appropriateness of the item is the Fisher item information function (Lord, 1980, equation 5-9) and the estimate of ability is maximum likelihood (Lord, 1980, equation 4-31), although other measures of the statistical properties of items (see for example, Chang, 1995) and other estimates of ability (see for example, Davey & Parshall, 1995) are possible. In the WDM used in this paper, the selection of the optimum next item is further moderated by the imposition of the extended Sympson & Hetter (1985) exposure control methodology (Stocking, 1992). In this methodology, exposure control parameters are developed through a series of simulations in a test design phase. These exposure control parameters are then used to restrict the frequency with which some items are administered even though they are selected initially as optimum by the WDM.

The Adaptive Tests

Item pools for adaptive measures of Quantitative, Verbal and Analytical Reasoning and Reading were obtained. The test design simulations for the first three measures are described in Eignor et al. (1993); those for the last measure are described in O'Neill, Folk, & Li (1993).

The item parameters for the Quantitative, Verbal and Analytical Reasoning pools were estimated from large samples of test-takers using the three parameter logistic item response model (3PL; Lord, 1980) and the computer program LOGIST (Wingersky, 1983). The item parameters for the Reading pool were estimated from smaller samples (500+) of test-takers using the 3PL model and the computer program BILOG (Mislevy & Bock, 1983).

All four tests used the WDM adaptive testing paradigm described earlier and used an estimated number right true score on a reference set of items as a raw adaptive test score. The test design simulations were conducted to establish the test lengths, exposure rates, constraint weights, and item pool sizes required to meet minimum desirable levels of estimated reliability (computed using Green, Bock, Humphreys, Linn & Reckase (1984), equation 6), and desirable conditional standard error of measurement (CSEM) curves. The simulations were conducted with reference to estimated distributions of true ability for the intended population, computed by the method of Mislevy (1984).

Table 5 contains specific information about each measure. The length of the adaptive test and the reference test used for scoring purposes are given in columns two and seven. The number of elements (discrete items, stimuli, and items belonging to stimuli) in each pool is given in column three. The number of explicit constraints on item and stimulus selection is given in the fourth column. All measures except the Reading measure had selection further

restricted by overlap groups. The number of sets of items administered in the adaptive test is given in column five, and the number of items represented by these sets of items is given in column six. The proportion of items in the adaptive test that are set-based range from a low of 14% for the Quantitative measure, to a high of 100%, that is, all items appear in sets, for the Reading measure. The final column gives the range of the reported score scale upon which some results will be reported.

Table 5: The Four Adaptive Tests

Test	CAT Length (items)	Number of Elements in Pool	Number of Constraints	Number of sets	Number of items in sets	Reference test length	Scaled score range
Quantitative	28	348	27	2	4	60	200-800
Verbal	30	381	38	3	8	76	200-800
Analytical	35	512	43	6	26	50	200-800
Reading	31	443	27	7	31	40	300-336

An Unrealistic Worst-Case Model of Revising Behavior

The approach taken in this paper to evaluating models for permitting test-taker revisions in adaptive testing is to assume a worst-case model of revising behavior. If a model for permitting revisions can be found that functions well under these conditions, then the same model will have better properties in actual adaptive test administrations to real test-takers. Under this worst-case model of test-taker revising behavior, it is assumed that, within the particular model for permitting item revisions:

1) All test-takers initially respond incorrectly to any item that will be subsequently revised. That is, all test-takers design the easiest possible test. This is unrealistic in that it implies perfect recognition of incorrect answer alternatives, even by test-takers with very low abilities.

2) All test-takers recognize that, given a choice, it is to their advantage to respond incorrectly initially to items presented earlier in the test or section rather than later in the test or section.

3) Revised responses are generated in accord with the psychometric model assumed to underlie examinee performance. For the models studied here, this was the 3PL.

Models for Permitting Revisions in CAT

Using the worst-case model of test-taker revising behavior and the WDM adaptive testing paradigm, three models for permitting item revisions in CAT were explored:

1) permit revisions to some number of previous answers,

2) permit revisions within separately timed test sections, and

3) permit revisions within sets of items belonging to a common stimulus.

All three models were studied using the actual adaptive test designs that have been used with real adaptive testing with live test-takers.

Model 1: Permit Revisions to Some Number of Items

Under this model, test-takers would be instructed in advance of testing that they will be permitted to revise answers to some (fixed) number of questions. Test-takers are sufficiently sophisticated that they understand

that to gain a high score it is in their best interests to attempt to answer items at the beginning of the test incorrectly, and then return to these items at the end of the test and revise all of their incorrect answers to correct answers, to the best of their ability.

Method:

Explorations of this model of permitting revisions used the Quantitative Reasoning adaptive test described earlier. The adaptive test simulations were conducted with uniform distributions of simulated examinees (simulees) across (nearly) equally spaced values on the score reporting metric, starting from about the chance score level and ending close to the top of the range. This results in values on the θ metric that are unequally spaced. In addition, to facilitate unconditional comparisons, a particular target population ability distribution was established. The target population ability distribution was estimated for these same (nearly) equally spaced values on the score reporting metric, using the method of Mislevy (1984) and a sample of over 6000 real test-takers who took an exemplar form of the linear version of this test. Conditional results were then weighted to reflect results for this target distribution.

The simulation experiment was performed four times: simulees could revise two items, seven items (one quarter of the test), 14 items (half of the test), and all 28 items in the test. It is this latter condition that reflects the Wainer (1992) worse-case speculation.

Model 1 Results:

The results of this model of permitting revisions in terms of the measurement properties of the resultant adaptive tests are shown in Figure 1a. The estimated distribution of true ability for the typical population is shown

as a histogram of proportional frequencies, with values to be read from the right vertical axis. The conditional standard errors of measurement at each true score level for various revision conditions are shown as curves with values to be read from the left vertical axis. Legend labels indicate the estimated reliability of the adaptive test under the various revision conditions.

The conditional standard error of measurement curves for two scorings of the conventional linear reference test -- observed number correct (lower) and estimated number correct true score (higher) are plotted as smooth solid curves. These curves were used for decision purposes in determining the final test design. The uneven solid curve is the CSEM curve for the adaptive test at the end of the test design simulations (labeled 'no revisions'). It was judged to be satisfactorily close to the CSEM curves for the reference test.

For true scores below 25, the number of items revised does not have much impact on the CSEM. This is because even with revising all items in the adaptive test, simulees at these low levels do not produce tests that are so easy that their observed test scores are not good approximations to their true scores. At the highest true score level, the same phenomenon also appears, but to a lesser extent, because these able simulees get nearly all of the items in any adaptive test correct and have no need to design an especially easy adaptive test.

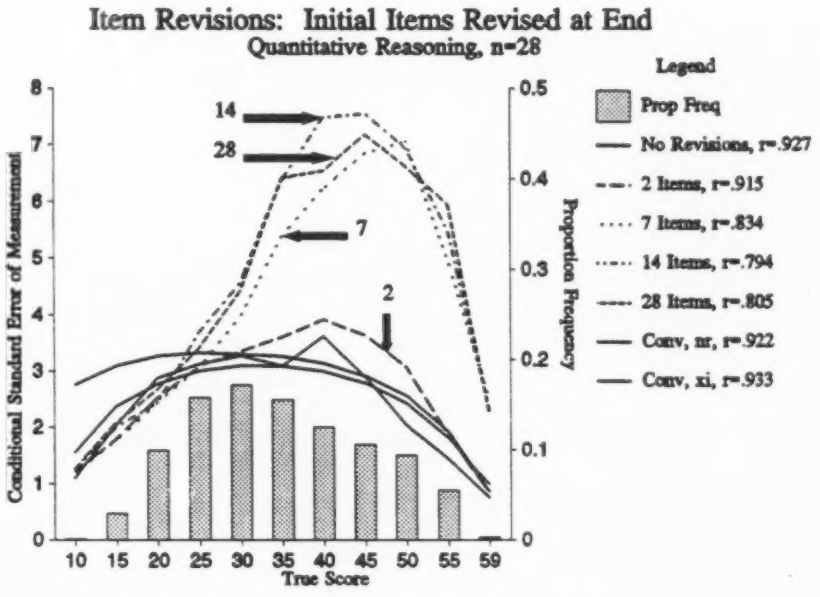

Figure 1a: Model 1, permit revision of some number of items at end of test.

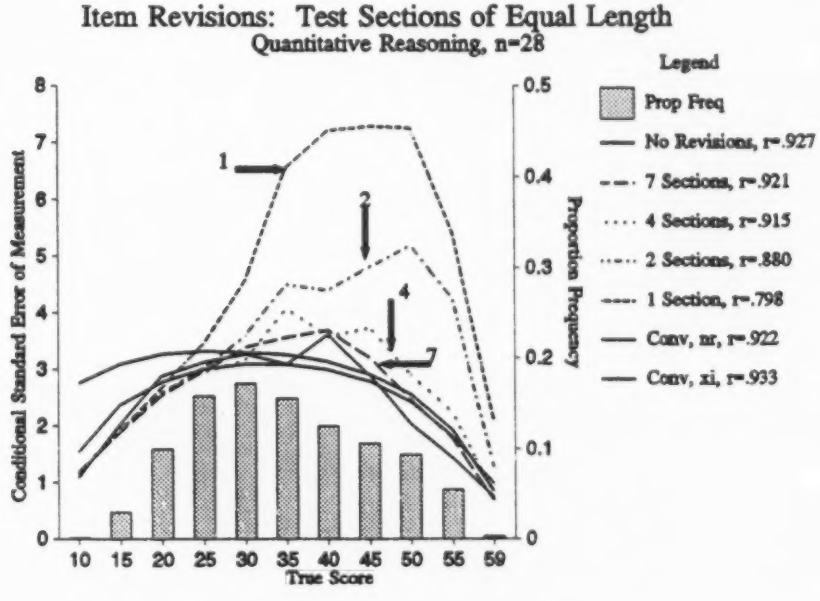

Figure 1b: Model 2, permit revisions in sections.

For middle true score levels, the number of items revised has a discernable impact. Even with only two items subject to revision, the CSEM is increased over the no revised items condition, and the estimated reliability is lowered. With seven revised items, the CSEM has more than doubled for some true score levels, while the estimated reliability has been substantially reduced; the results are even more extreme for 14 and 28 items. Under the assumed worst-case model of examinee revising behavior, this model of permitting test-taker revisions results in a significant degradation of the measurement properties of the test.

Results of this model of permitting test-taker revisions in terms of differences in the resultant distributions of test scores is shown in Figure 2a. For each condition, that is, each number of revised items, the difference in test scores from the no revision condition was computed (revised condition minus no revision condition) for the 25th, 50th, 75th, 90th, 95th, 98th, and 99th percentiles of the distribution of test scores for a random sample of 1000 simulees from a typical population of test-takers. The no revision simulation condition was then repeated (with a different random number seed), and the same difference scores were computed for this condition also. These difference scores represent what one might expect upon retesting from the same item pool. The difference between these two difference scores represents the changes due to the revision condition over and above what one might see upon a simple retesting with the same pool and no revisions. The results are reported in terms of the reported scaled score metric that ranges from a low of 200 to a high of 800.

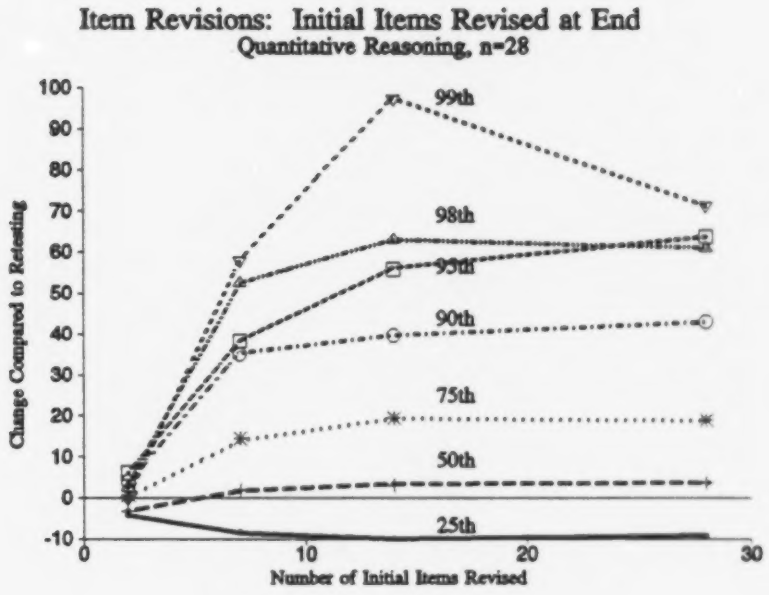

Figure 2a: Model 1, change in percentile values compared to retesting.

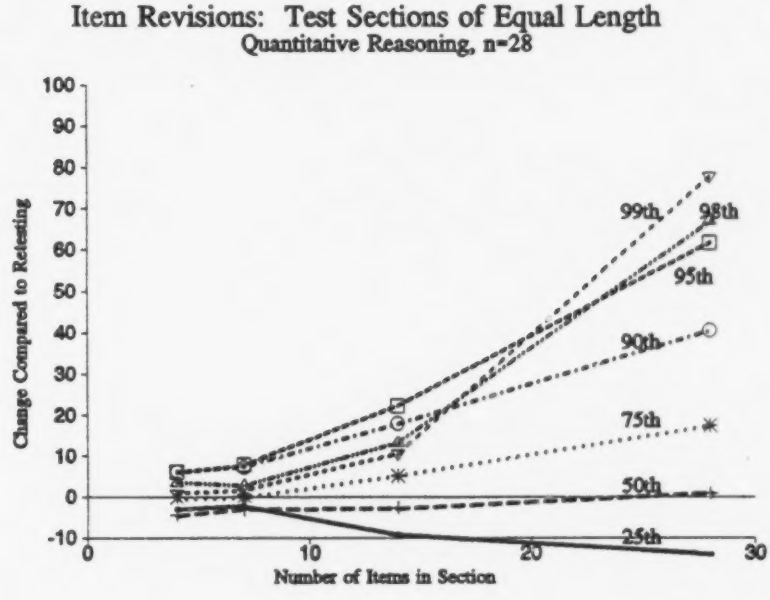

Figure 2b: Model 2, change in percentile values compared to retesting.

Regardless of the number of item revisions permitted, the 50th percentile of the score distribution obtained with revisions is within a few scaled score points (less than five) of what one might expect upon simple retesting with the same item pool. With only two items revised, all of the percentile points are likewise similar to what one might expect upon retesting.

However, even with as few as seven revised items, the 95th percentile of the distribution of resultant scores is between 30 and 40 scaled score points higher than what might be seen upon retesting, as is the 90th percentile for the 14 and 28-item conditions. If all 28 items are revised, the 95th, 98th, and 99th percentiles of the resultant distribution of test scores are more than 60 scaled score points higher than might be obtained from retesting. Differences of this magnitude in reported score distributions due to permitting test-taker revisions are unlikely to be ignorable.

Model 2: Permit Revisions Within Separately Timed Sections

Under this model, the adaptive test is divided into separately timed sections and test-takers are informed in advance of testing that they will be permitted to revise answers to questions only within a section, identical to current practice on linear paper-and-pencil tests. In the environment of adaptive testing, separately timed sections are formed by grouping together neighboring items as they are administered to the test-taker. Thus the first x items constitute the first section, items appearing in position x+1 to 2x constitute the second section, items appearing in positions 2x+1 to 3x constitute the third section, and so forth. The actual items appearing in a section would, of course, differ by test-taker, but the number of items would

not. An alternative name for this type of construction might be 'block review', where the contents of a block are not fixed in advance.

The advantage of this model over the previous model is twofold. First, it can be used to permit even more test-taker control over item revisions, that is, all items in the test may be revised, regardless of the number of sections into which the test is divided. Second, it simultaneously restricts test-taker control over the actual items presented because revised responses from previous sections influence the selection of items in subsequent sections. This is in contrast to the previous model, where revised responses have no impact on item selection since revisions do not take place until after the last item has been selected. Thus, if a test-taker has used the strategy of designing an inappropriately easy section on which they receive a very high score, the item selection algorithm automatically compensates for this by selecting harder items in the next section.

Method:

As before, the adaptive test simulations were conducted for the Quantitative Reasoning pool with uniform distributions of simulees across (nearly) equally spaced values on the score reporting metric and the results weighted to reflect results for the target population. The experiment was performed four times: the adaptive test was considered to be seven sections of four items each, four sections of seven items each, two sections of 14 items each, and one section of 28 items. This last condition constitutes a replication of the 28-item full review condition for Model 1.

Model 2 Results:

The results of this model of permitting revisions in terms of the measurement properties of the resultant adaptive tests are shown in Figure 1b,

where the axes have the same meanings as in Figure 1a. At low ability levels, the results for adaptive tests divided into sections are similar to those shown in Figure 1a, and for the same reasons, namely simulees at such low levels cannot design tests that are so easy that their observed scores are not good approximations to their true scores.

At middle and high ability levels, tests divided into more sections with fewer items (seven sections with four items each and four sections with seven items each) have CSEMs and estimated reliabilities close to those obtained when no revisions are permitted. Even a two section test (14 items in each section), although not acceptable from a measurement perspective, shows substantial improvement over the 14-item review condition for Model 1 in which revisions are not incorporated in subsequent item selections. The smaller the test section, the more rapidly the effects of item revisions are incorporated into the item selection algorithm, thus mitigating test-taker attempts to design inappropriately easy tests.

Similar encouraging results for the differences in distributions of reported scores are seen in Figure 2b. The largest difference over retesting from the same item pool when the test is considered a two-section (14-item each) test is at the 95th percentile, but this difference is only a little more than 20 scaled score points. Dividing the adaptive test into either four or seven sections produces changes, when compared to retesting, at all percentiles that are less than 10 scaled score points.

Model 3: Permit Revisions Within Sets of Items Belonging to a Common Stimulus

Under this model, test-takers are permitted to revise answers to questions only within a set of items that are associated with common stimulus

material. Revisions to answers for items not associated with a common stimulus are not permitted. This model has some of the features of the previous model. Blocks are now formed naturally by association with related stimulus material, rather than artificially as in Model 2 where blocks were formed from arbitrary and perhaps unrelated items. Because of this natural formulation by sets, the number of items available for revision varies from set to set. Ultimately, if the entire adaptive test consists of sets of items constructed in this fashion, this model may be viewed as a more general case of the previous model in that, although all items may be revised, the number of items that may be revised at any one point in time is variable and, typically, small.

The advantage of this model over the previous model is that the formation of groups of items that may be revised may be more consonant with cognitive processing demands. Certainly the linear item revision data discussed earlier suggests that this may be the case. This model also retains the advantage of the previous model of insuring that revised responses impact subsequent choices of items by the item selection algorithm. The disadvantage of this model over the previous model is that there is no review of discrete items that are not associated with a common stimulus. Thus, the test-taker has some, but not complete, control over the order of item responses.

Method:

All four adaptive tests describe earlier were used to explore this model of permitting revisions in adaptive tests. This was necessary in order to study tests with different numbers of sets and items belonging in sets, as given in columns five and six in Table 5. As before, the adaptive test simulations were conducted with uniform distributions of simulees across

(nearly) equally spaced values on the (raw) score reporting metric and the results weighted to reflect results for the target population. A single experiment was performed for each item pool.

Model 3 Results:

Figure 3 shows the results of this model in terms of the resultant measurement properties of the adaptive tests. All panels have the same values on the right vertical axis for the reading of proportional frequencies. The two panels in the first row are for the Quantitative and Verbal Reasoning measures and have the same values on the left vertical axis for reading the CSEMs. The two panels in the second row are for the Analytical Reasoning and Reading measures and have the same values on the left vertical axis, which differ from those in the first row.

On all panels, the CSEM curves for the reference set of items and for the adaptive tests at the end of the test design phase are drawn as solid curves. The CSEMs for these adaptive tests represent the no revision condition; the CSEM curves for the reference set of items represent the standard of comparison to which the adaptive test designs were held in deciding when to end the test design phase. On each panel there is a dotted line that is the result of the extreme Model 1 condition -- revise all items at the end of the test. This represents the Wainer worst-case scenario and is provided for comparison. The thick dashed line on each panel represents the results for the current model -- revisions are permitted only within sets of items.

For all adaptive tests, permitting item revisions within sets only is as satisfactory, from a measurement perspective, as prohibiting revisions entirely. This occurs for two different reasons. For measures with few set-

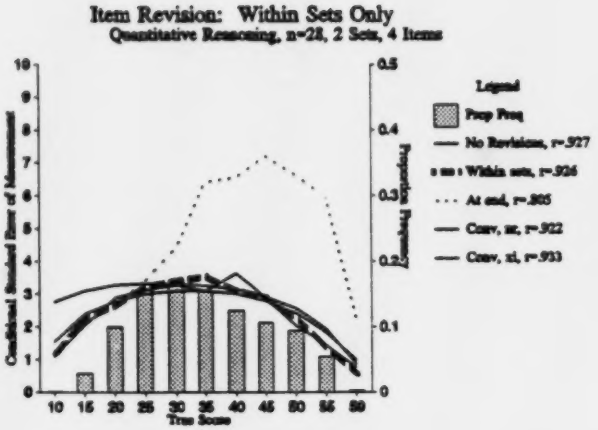

Figure 3a: Model 3, Quantitative Reasoning

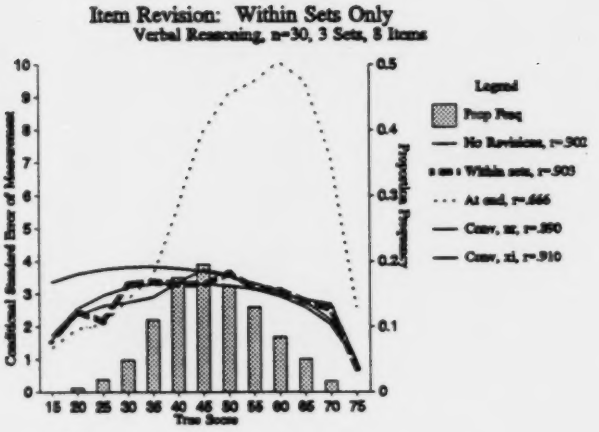

Figure 3b: Model 3, Verbal Reasoning

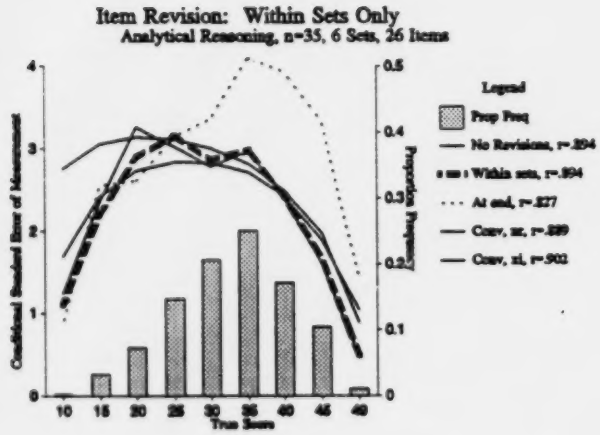

Figure 3c: Model 3, Analytical Reasoning

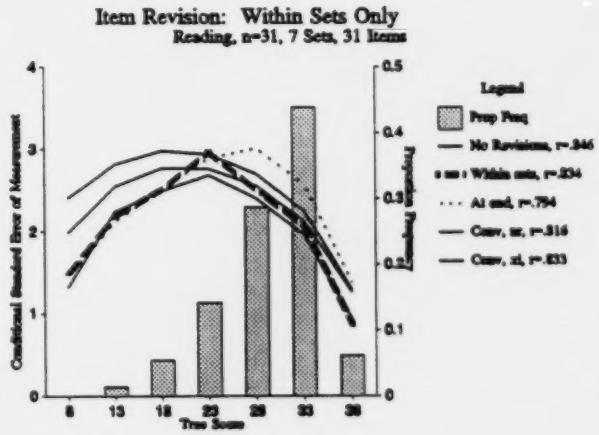

Figure 3d: Model 3, Reading

based items, such as the Quantitative and Verbal measures, few items get revised therefore impact is minimal and these few revisions are reflected in the choice of subsequent items. For measures with more set-based items, such as the Analytical and Reading measures, more items are available for revision, but these revisions take place within set restrictions and also effect subsequent item selection.

For the Reading measure, all items are revised since all items are in sets. For this measure, the small difference between the current model and the most extreme Model 1 case is due to the fact that it is difficult, in an all set-based test, to employ the Wainer suggestion effectively if items in sets tend to be more heterogeneous than desirable, as they are in this case. The difference between these two conditions would be greater if sets of items were more homogeneous in difficulty, thus allowing the Wainer strategy to be more effectively employed.

Differences in distributions of reported scaled scores, when compared to retesting, are displayed in Figure 4 for Model 3 and Model 1 results. The tests are displayed in the same positions as in Figure 3. In contrast with Figure 2, the horizontal axis in Figure 4 serves to artificially locate the two models being compared for each measure. The left vertical axis is the same for the three measures with the same 200 to 800 scaled score metric; it differs for the Reading measure that has a scaled score range of 300 to 336.

For the Quantitative, Verbal, and Analytical measures, all percentiles computed differ from those expected on retesting by less than 10 scaled score points when revisions are permitted only within sets of items. This is true whether there are few sets, as in the Quantitative measure, or many sets, as in the Analytical measure. For the Reading measure, all differences in

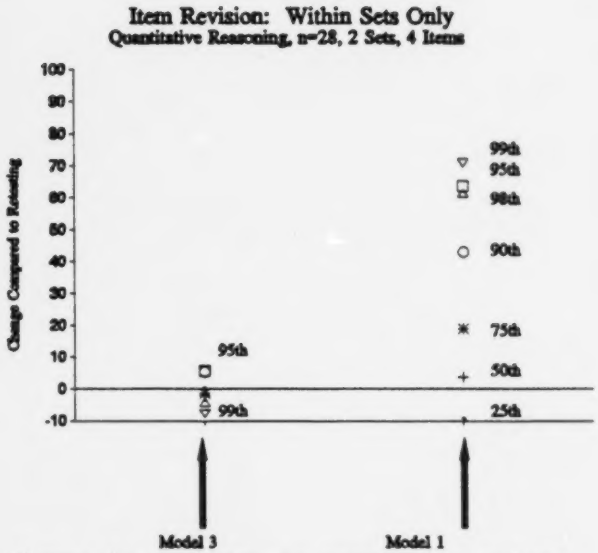

Figure 4a: Change in percentile values compared to retesting.

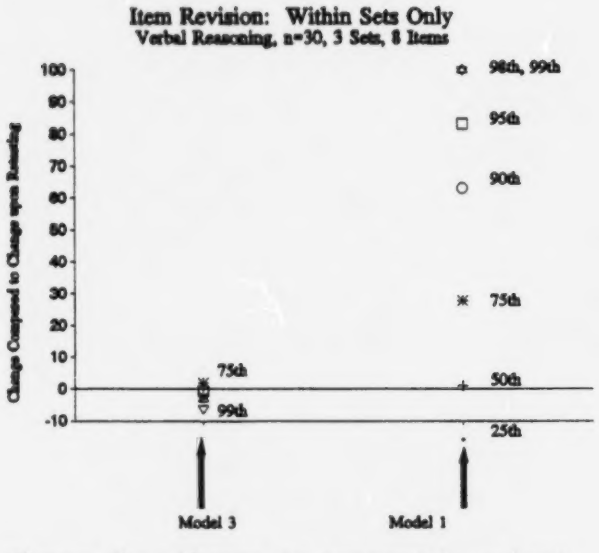

Figure 4b: Change in percentile values compared to retesting.

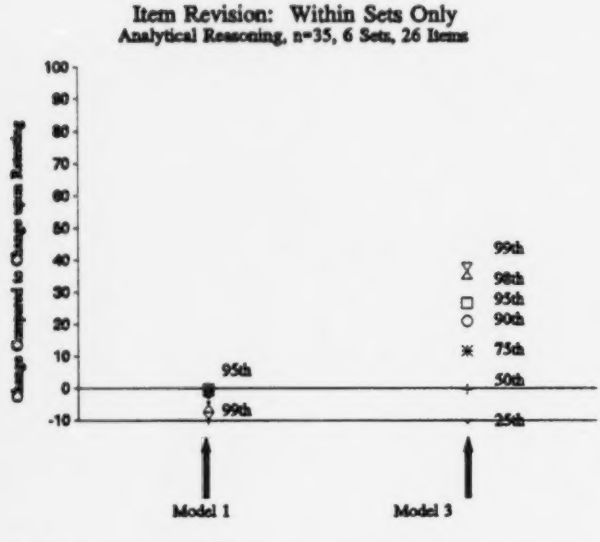

Figure 4c: Change in percentile values compared to retesting.

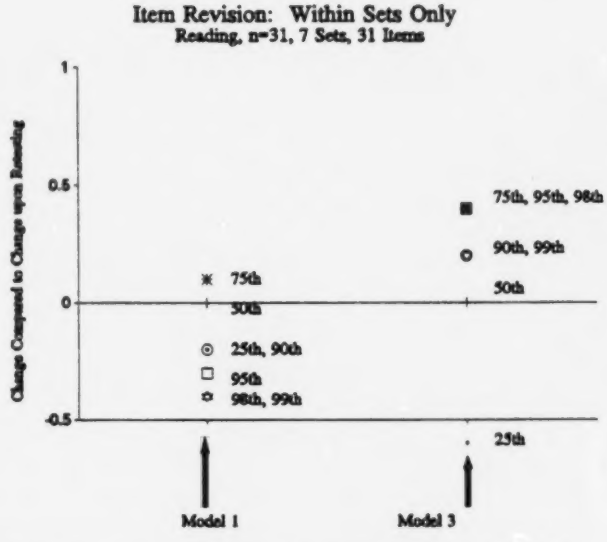

Figure 4d: Change in percentile values compared to retesting.

percentiles lie within about a half point on the reported score metric of what might be expected from retesting, for both the Model 3 and Model 1 conditions. However, most of the differences from retesting are negative for the Model 3 condition and most of the differences are positive for the Model 1 condition.

Discussion and Conclusions

In linear testing, where items are chosen in advance by test designers and every test-taker receives the same (or parallel) sets of items, the order in which final answers are supplied to questions is under the control of the test-taker. In adaptive testing, where new questions are chosen as current questions are answered, unthinking attempts to provide the same feature can result in test-taker control over which items are actually presented in addition to the order of responses. Neither the study of revising behavior in linear testing, nor previous studies of revisions in CAT, inform the situation for the environment envisioned by Wainer (1992) in which test-takers are well-informed of strategies that may maximize test scores in an adaptive testing context. If these strategies are used by all test-takers, the adaptive test becomes worthless from the perspective of the test-score user but will be fair to all test-takers; if such strategies are used by only some test-takers, the adaptive test becomes unfair to those test-takers who do not employ them.

Using the unrealistic worst-case model of test-taker revising behavior described earlier, three models of permitting item revisions in the CAT environment were explored. The first model most closely mirrors the environment envisioned by Wainer in that revisions are permitted to all or some test-taker chosen subset of items, and the best test-taker strategy is to intentionally select incorrect responses to items at the beginning of the test

and make permitted revisions only after all items have been administered. In this fashion, intentionally incorrect responses impact the greatest number of subsequent item selections and revised responses to those items do not influence the choice of subsequent items. If more than a few revisions are permitted, the consequence of this model of permitting revisions is to seriously impair the measurement properties of the test.

The other two models studied attempt to restrict revisions to blocks of items, thus forcing revised responses to influence subsequent item selections. In Model 2, an adaptive test is considered to consist of separately timed sections or blocks of items of fixed length but variable content. If the number of sections is relatively large, this model provides effective control over test-taker strategies that impair the measurement efficiency of adaptive testing under Model 1, while also allowing the review of all items administered.

Model 3 permits revisions only within sets of items associated with common stimulus material. This is consonant with the suggestion from the study of linear revising behavior that at least some test-takers prefer to respond to such items only after all items in the set have been considered. Thus revisions are permitted in blocks of arbitrary length and position throughout a test and the number and location of such sets depends upon the particular adaptive test design. This model, when applied to a variety of adaptive tests with different numbers and sizes of sets, also provides effective control over strategies that impaired the measurement efficiency under Model 1, while at the same time permitting revisions that may be more in accord with cognitive processing demands than those of Model 2. A possible disadvantage of this model is that revisions to discrete items are not

permitted. However, it could be argued that permitting revisions to discrete items in adaptive testing (as opposed to linear testing) is not necessary since in adaptive testing test-takers are not presented with items that are much too hard for them.

It seems possible, then, to transfer some of the features of the more familiar linear testing environment to the less familiar adaptive testing environment, as long as this transfer takes into account fundamental differences between the two contexts. Moreover, it is possible to make this transfer in a fashion that preserves important cognitive processing styles developed by test-takers, while eliminating those that are not appropriate to this new environment.

References

Eignor, D. R., Way, W. D., Stocking, M. L., & Steffen, M. (1993). Case studies in computer adaptive test design through simulations. (Research Report 93-56). Princeton, NJ: Educational Testing Service.

Chang, H. (1995). A global information approach to computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education. April, 1995, San Francisco, CA.

Davey, T., & Parshall, C. G. (1995). New algorithms for item selection and exposure control with computerized adaptive testing. Paper presented at the annual meeting of the American Educational Research Association. April, 1995, San Francisco, CA.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 347-360.

Jacobson, R. L. (1993, September 13). New computer technique seen producing a revolution in testing. The Chronicle of Higher Education, p A22.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1, 95-100.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. Applied Psychological Measurement, 16, 33-40.

Mills, C. N., & Stocking, M. L. (1995). Practical issues in large-scale computerized adaptive testing. (Research Report 95-23). Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1984). Estimating latent distributions. Psychometrika, 49, 359-381.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. Psychometrika, 51, 177-195.

Mislevy, R. J., & Bock, R. D. (1983). BILOG: Item and test scoring with binary logist models [computer program]. Mooresville, IN: Scientific Software.

O'Neill, K., Folk, V., & Li, M.-Y. (1993). Report on the pretest calibration study for the computer-based academic skills assessments of The Praxis Series: Professional Assessments for Beginning Teachers (TM). Princeton, NJ: Educational Testing Service.

Schaeffer, G., Steffen, M., & Golub-Smith, M. (1993) Introduction of a computer adaptive GRE general test (Research Report (93-57). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1992). Controlling item exposure rates in a realistic adaptive testing paradigm (Research Report 93-2). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement, 17(3), 277-292.

Stone, G. E., & Lunz, M. E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. Applied Measurement in Education, 7(3), 211-222.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. Applied Psychological Measurement, 17, 151-166.

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing, as described in Wainer, et al., (1990).

Wainer, H. (1992). Some practical considerations when converting a linearly administered test to an adaptive format. (Research Report 92-13). Princeton, NJ: Educational Testing Service.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J.,Steinberg, L., & Thissen, D. (1990). Computerized Adaptive Testing: A Primer. Hillsdale, NJ: Lawrence Erlbaum Associates.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.

Wingersky, M. S. (1983) LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.